CLOUD APPLICATION DEVELOPMENT (GROUP 1)

PHASE 5 : ASSIGNMENT NOTEBOOK SUBMISSION

NAME :Tamil Selvi.P

PROJECT TITLE : Big Data Analysis with IBM Cloud Databases[05]

EMAIL ID : tamiljesus2004@gmail.com

GUIDED BY: MRS.J.HEMALATHA

SPOC NAME : MR. P. VIGNESH

GIT HUB REPOSITORY URL :

https://github.com/TamilSelvip-2004/Bigdata-analysis/settings/keys

BIG DATA ANALYSIS

INTRODUCTION :

***Big data** primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many entries offer greater statistical power, while data with higher complexity may lead to a higher false discovery rate . Big data analytics is the process of collecting, examining, and analysing large amounts of data to discover market trends, insights, and patterns that can help companies make better business decisions. Two conspicuous examples are Amazon Prime, which uses Big Data analytics to recommend programming for individual users, and Spotify, which does the same to offer personalized music suggestions.*

Problem Definition:

The project involves delving into big data analysis using IBM Cloud Databases. The objective is to extract valuable insights from extensive datasets, ranging from climate trends to social patterns. The project includes designing the analysis process, setting up IBM Cloud Databases, performing data analysis, and visualizing the results for business intelligence.

Design Thinking:

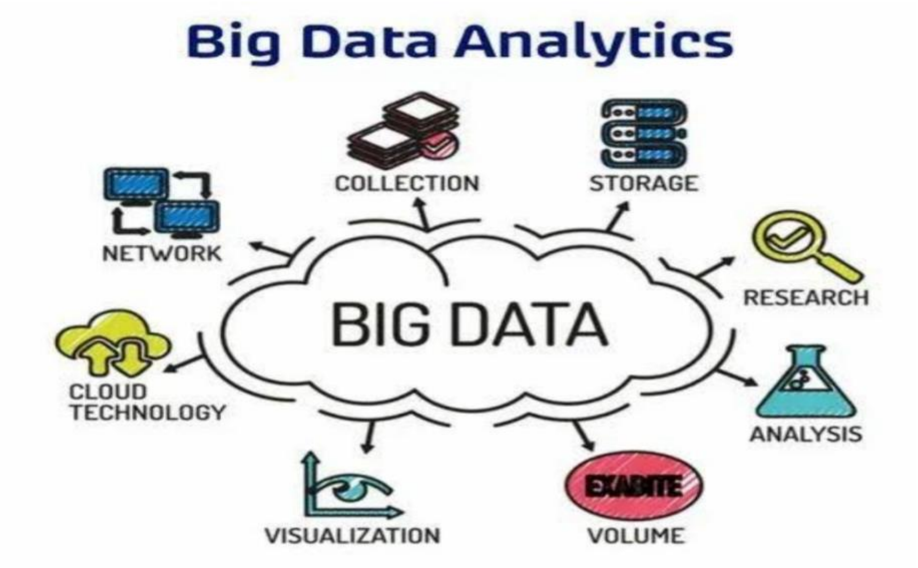
1. *Data Selection: Identify the datasets to be analyzed, such as climate data or social media trends.*
2. *Database Setup: Set up IBM Cloud Databases for storing and managing large datasets.*
3. *Data Exploration: Develop queries and scripts to explore the datasets, extract relevant information, and identify patterns.*
4. Analysis Techniques: Apply appropriate analysis techniques, such as statistical analysis or machine learning, to uncover insights
5. Visualization: Design visualizations to present the analysis results in an understandable and impactful manner.
6. Business Insights: Interpret the analysis findings to derive valuable business intelligence and actionable recommendations.

What is Big Data?

Big data defined

What exactly is big data?

- ➢ The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.
- ➢ Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can’t manage them.



purpose of big data :

With big data, you can analyze and assess production, customer feedback and returns, and other factors to reduce outages and anticipate future demands. Big data can also be used to improve decision-making in line with current market demand.

How big data works ?

Big data gives you new insights that open up new opportunities and business models. Getting started involves three key actions:

1. Integrate

Big data brings together data from many disparate sources and applications. Traditional data integration mechanisms, such as extract, transform, and load (ETL) generally aren't up to the task. It requires new strategies and technologies to analyze big data sets at terabyte, or even petabyte, scale.

During integration, you need to bring in the data, process it, and make sure it’s formatted and available in a form that your business analysts can get started with.

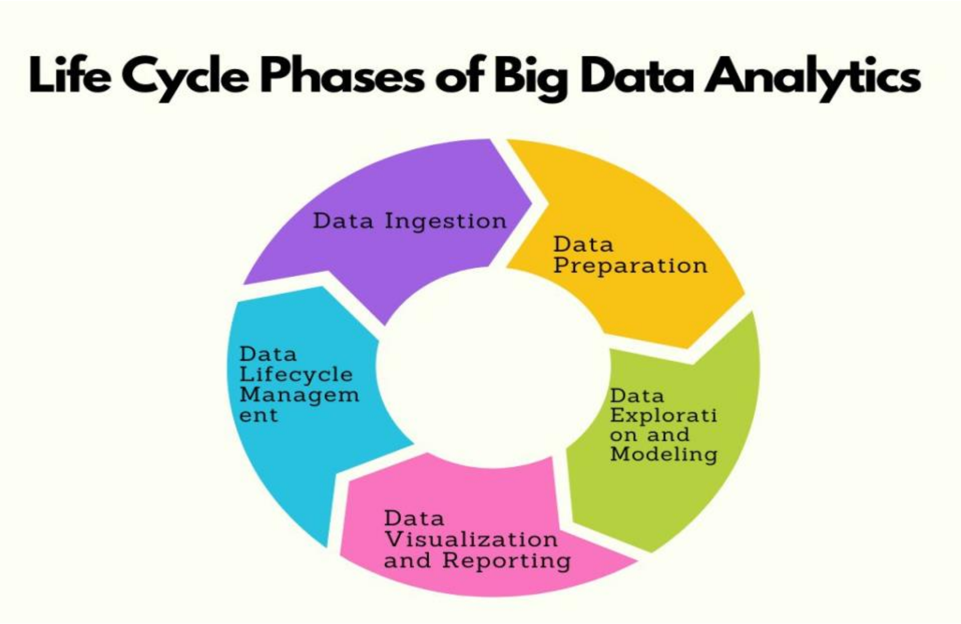
2. Manage

Big data requires storage. Your storage solution can be in the cloud, on premises, or both. You can store your data in any form you want and bring your desired processing requirements and necessary process engines to those data sets on an on-demand basis. Many people choose their storage solution according to where their data is currently residing. The cloud is gradually gaining popularity because it supports your current compute requirements and enables you to spin up resources as needed.

3. Analyze

Your investment in big data pays off when you analyze and act on your data. Get new clarity with a visual analysis of your varied data sets. Explore the data further to make new discoveries. Share your findings with others. Build data models with machine learning and artificial intelligence. Put your data to work.

.


Life Cycle Phases of Big Data Analytics
Data Ingestion
Data Preparation
Data Explorati on and Modeling
Data Visualization and Reporting
Data Lifecycle Managem ent

Here is a diagram representing this process:

Life Cycle Phases of Big Data Analytics

The following are the phases in the life cycle of big data analytics in brief: 3

● **Data Ingestion:** This is the process of collecting, extracting, and loading data from various sources into a centralized data repository.

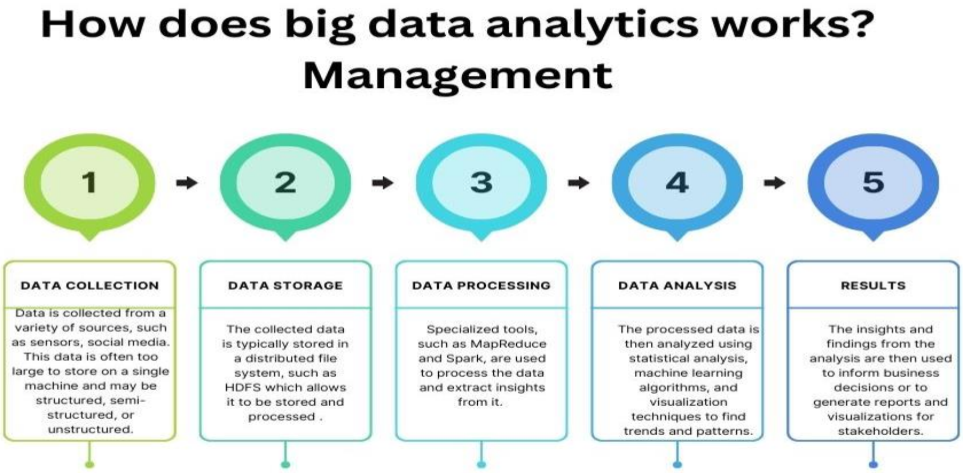
● **Data Preparation:** This is the cleaning, transforming, and preparing of data for analysis.

● **Data Exploration** and Modeling: This is the process of using various analytical techniques and tools to uncover patterns and insights in the data.

● **Data Visualization and Reporting:** This is the process of using visual aids to communicate the findings from the data analysis.

● **Data Lifecycle Management:** This is the process of managing the data throughout its lifecycle, from ingestion to visualization and reporting. The data can be collected from various sources like IoT devices, Social Media, Images and

Videos, Applications, and RFID and can be stored and integrated from diverse sources like NoSQL Data Bases, Data Warehouses, Data Lakes.



How does big data analytics works?
Management

1 DATA COLLECTION: Data is collected from a variety of sources, such as sensors, social media. This data is often too large to store on a single machine and may be structured, semi-structured, or unstructured.

2 DATA STORAGE: The collected data is typically stored in a distributed file system, such as HDFS which allows it to be stored and processed .

3 DATA PROCESSING: Specialized tools, such as MapReduce and Spark, are used to process the data and extract insights from it.

4 DATA ANALYSIS: The processed data is then analyzed using statistical analysis, machine learning algorithms, and visualization techniques to find trends and patterns.

5 RESULTS: The insights and findings from the analysis are then used to inform business decisions or to generate reports and visualizations for stakeholders.

Consider incorporating advanced machine learning algorithms for predictive analysis or anomaly detection in the big data.

Ensemble Learning: Random Forest is an ensemble of decision trees, which combines multiple models for improved accuracy.

Decision Trees: Each tree in the forest is constructed based on different subsets of data, reducing overfitting.

Importance Scores: It provides feature importance scores, helping you identify which features are most influential.

Gradient Boosting: XGBoost, LightGBM, CatBoost: These are gradient boosting libraries known for their speed and performance.

Boosting Process: Boosting iteratively improves the model’s performance by focusing on data points that were previously misclassified.

Regularization: These algorithms offer regularization techniques to prevent overfitting.

Neural Networks: Convolutional Neural Networks (CNNs): Ideal for image recognition tasks, they use convolutional layers to detect patterns in images.

Recurrent Neural Networks (RNNs): Suitable for sequential data, they capture temporal dependencies.

Deep Learning: Deep neural networks with many layers can learn intricate patterns but require substantial data and computational resources.

Development Part 1: In this section begin building your project by loading and preprocessing the dataset.

Data Collection:

- Methods for gathering big data.
- Sources of big data (e.g., social media, IoT d
- evices, sensors).

Data Preprocessing:

- Data cleaning and quality assurance.
- Handling missing data in big datasets.
- Data transformation and normalization.

Data Storage:

- Choosing the right data storage solutions (e.g., Hadoop HDFS, cloud storage).
- Managing data formats (e.g., JSON, Parquet, Avro).

Data Integration:

- Combining data from multiple sources.
- Data integration challenges in big data projects.

Data Sampling:

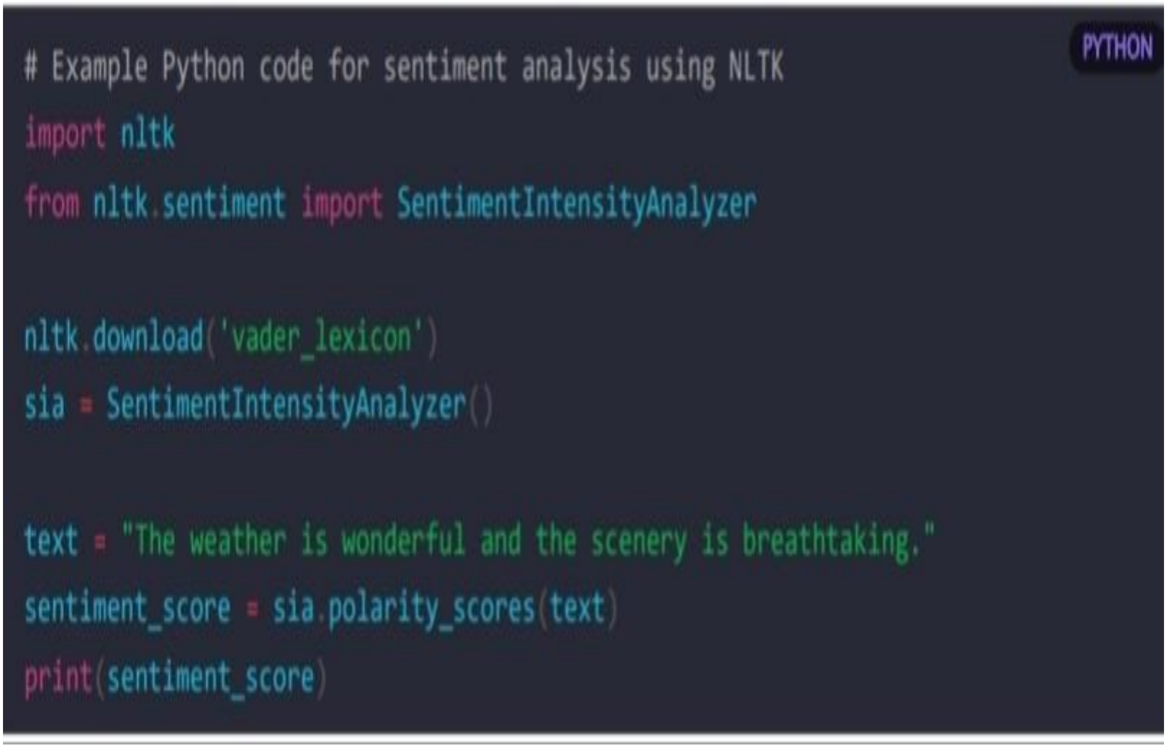
- Strategies for reducing the size of large datasets.
- Random vs. stratified sampling.

Building the big data analysis solution by applying advanced analysis techniques.

Advanced analysis techniques:

#Apply more complex analysis techniques, such as machine learning algorithms, time series Analysis, or sentiment analysis, depending on the dataset and objectives.

#There are many algorithms like random forest, gradient descent, etc… .. …



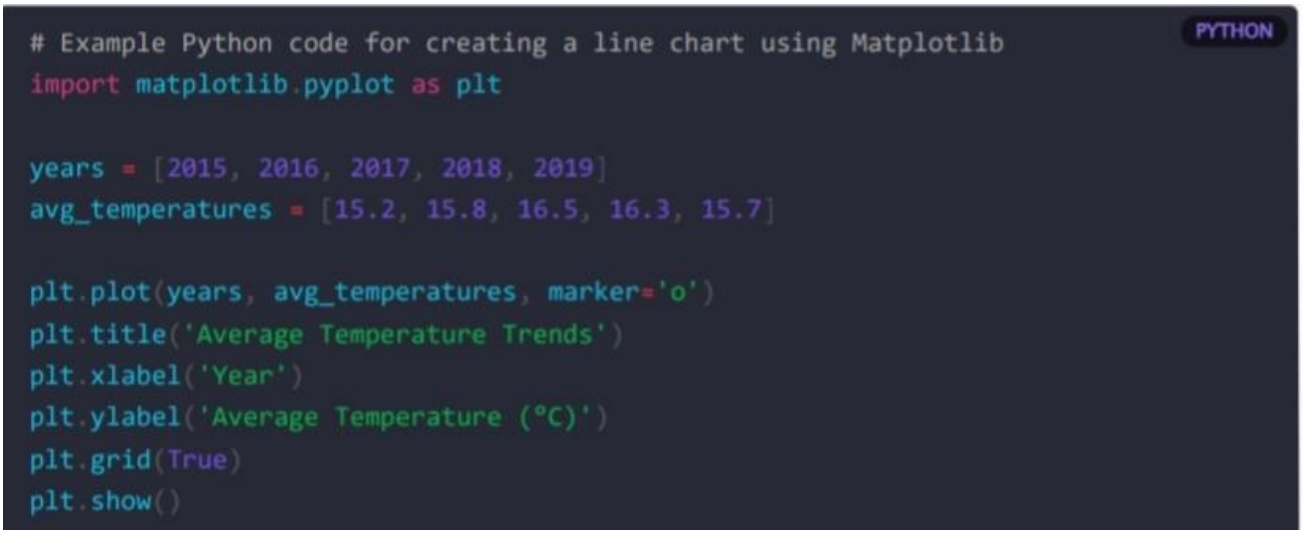
```python
# Example Python code for sentiment analysis using NLTK
import nltk
from nltk.sentiment import SentimentIntensityAnalyzer

nltk.download('vader_lexicon')
sia = SentimentIntensityAnalyzer()

text = "The weather is wonderful and the scenery is breathtaking."
sentiment_score = sia.polarity_scores(text)
print(sentiment_score)
```

Visualization :

Create visualizations to showcase the analysis results. Use tools like Matplotlib, Plotly, or IBM Watson Studio for creating graphs and charts.



```python
# Example Python code for creating a line chart using Matplotlib
import matplotlib.pyplot as plt

years = [2015, 2016, 2017, 2018, 2019]
avg_temperatures = [15.2, 15.8, 16.5, 16.3, 15.7]

plt.plot(years, avg_temperatures, marker='o')
plt.title('Average Temperature Trends')
plt.xlabel('Year')
plt.ylabel('Average Temperature (°C)')
plt.grid(True)
plt.show()
```

CLIMATIC CAT-07

Analysing climatic data to identify trends and patterns in big data is a complex and vital task in climate science. This process involves collecting and processing vast amounts of meteorological information, such as temperature, precipitation, wind patterns, and more, often gathered from various sources like weather stations, satellites, and climate models.

The primary goals of this analysis are to:

Identify Long-Term Trends: By examining historical data, scientists can determine whether certain climatic variables, like temperature, are consistently rising or falling over elxtended periods. This helps in assessing the presence of global warming or other significant climate shifts.

Detect Seasonal Patterns: Analyzing shorter-term data can reveal recurrent patterns related to seasons. This is critical for understanding phenomena like monsoons, hurricanes, and the changing of the seasons.

Extreme Events Analysis: Identifying trends in extreme weather events, such as heatwaves, droughts, or intense storms, is essential for climate adaptation and disaster preparedness.

Correlation and Causation: Researchers also investigate how various climatic factors interact with one another. For example, does rising sea surface temperature correlate with increased hurricane intensity?

Model Validation: Climate models are constantly refined and improved through data analysis. Comparing model predictions to real-world data helps ensure the models accurately simulate climate systems.

Regional Variations: Understanding how climate change affects specific regions is crucial. This involves analyzing localized climatic data and its implications for agriculture, water resources, and ecosystems

CONCLUSION :

- The fact is, big data analytics is not just a passing fad. It's a game changer that is only going to become more important in the years to come. Businesses that don't unlock the potential of big data analytics are going to be left behind.

- So what are you waiting for? Start exploring the possibilities of big data analytics today and see how you can improve your business performance. By now, you should know what big data analysts do, the skills they need, and the career opportunities available to them. The world of big data is growing rapidly, and analysts who are able to harness the power of big data analytics will be in high demand. So what are you waiting for? Start learning today and see if a career in big data analytics is right

GitHub Repository URL: https:/github.com/TamilSelvip-2004/

ps://gBigdata-analysis/settings/keys